

U.S. Department of Education  
November 2016

---

# Early Implementation Findings From a Study of Teacher and Principal Performance Measurement and Feedback: Year 1 Report Executive Summary

---

Andrew J. Wayne  
Michael S. Garett  
Seth Brown  
Jordan Rickles  
Mengli Song  
David Manzeske  
American Institutes for Research

Melanie Ali  
*Project Officer*  
Institute of Education Sciences

This page has been left blank for double-sided copying.

---

# **Early Implementation Findings From a Study of Teacher and Principal Performance Measurement and Feedback: Year 1 Report Executive Summary**

---

**November 2016**

**Andrew J. Wayne**  
**Michael S. Gare**  
**Seth Brown**  
**Jordan Rickles**  
**Mengli Song**  
**David Manzeske**  
American Institutes for Research

**Melanie Ali**  
*Project Officer*  
Institute of Education Sciences

**NCEE 2017-4003**  
**U.S. DEPARTMENT OF EDUCATION**

This page has been left blank for double-sided copying.

**U.S. Department of Education**

John King  
*Secretary*

**Institute of Education Sciences**

Ruth Curran Neild  
*Deputy Director for Policy and Research*  
*Delegated Duties of the Director*

**National Center for Education Evaluation and Regional Assistance**

Joy Lesnick  
*Acting Commissioner*

November 2016

This report was prepared for the Institute of Education Sciences under Contract ED-IES-11-C-0066. The project officer is Melanie Ali in the National Center for Education Evaluation and Regional Assistance.

IES evaluation reports present objective information on the conditions of implementation and impacts of the programs being evaluated. IES evaluation reports do not include conclusions or recommendations or views with regard to actions policymakers or practitioners should take in light of the findings in the report.

This report is in the public domain. Authorization to reproduce it in whole or in part is granted. While permission to reprint this publication is not necessary, the citation should be:

Wayne, A. J., Garet, M. S., Brown, S., Rickles, J., Song, M., Manzeske, D., (2016). *Early Implementation Findings From a Study of Teacher and Principal Performance Measurement and Feedback: Year 1 Report, Executive Summary* (NCEE 2017-4003). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

This report is available on the IES website at <http://ies.ed.gov/ncee>.

**Alternate Formats:** Upon request, this report is available in alternate formats, such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-9895 or 202-205-8113.

This page has been left blank for double-sided copying.

## Acknowledgments

This study was a collaborative effort and involved a diverse group of partners. We were fortunate to have had the advice of our expert technical working group. Members included Thomas Cook, Northwestern University; Thomas Dee, Stanford University; Laura Goe, Educational Testing Service; Laura Hamilton, RAND; Daniel McCaffrey, Educational Testing Service; Catherine McClellan, Clowder Consulting; Jonah Rockoff, Columbia University; Carla Stevens, Houston Independent School District; John Tyler, Brown University; and Judy Wurtzel, Charles and Lynn Schusterman Foundation.

We would also like to thank all those who provided the teacher and principal performance feedback systems and training, including the organizations that supported the implementation of the Classroom Assessment Scoring System (the University of Virginia and Teachstone), the Framework for Teaching (the Danielson Group and Teachscape), and the VAL-ED (Discovery Education). We appreciate the willingness and commitment of the school district leaders, schools, principals, study-hired observers, and teachers to implement the intervention and data collection activities, which involved a significant amount of time and energy.

We are also grateful to the AIR staff who worked diligently to coordinate the study's performance feedback activities in participating districts: Rebecca Herman, Fran Stancavage, Matthew Clifford, Mariann Lemke, Susan Ward, Carmen Martinez, Muna Shami, Ben Kalina, Marlene Darwin, Carla Hulce, Nicole Christie, Debbie Davidson-Gibbs, Mark Garabaldi, Jessica Milton, Elaine Liebesman, Amy Potemski, Roshni Menon, Marian Eaton, Gur Hoshen, Zhongjie Sun, and Michele Cadigan. Additional AIR staff worked tirelessly on data collection: Dorothy Seidel, Lauren Staley, Cheryl Pruce, Sara Bardack, Makeda Amelga, and Lindsey Mitchell. For their efforts to identify the partner districts, we thank the recruitment leaders Anja Kurki and Rebecca Herman and the team of senior recruiters: Kirk Walters, James Taylor, Marlene Darwin, Nicholas Sorensen, Mark Garibaldi, Carmen Martinez, Nicol Christie, Kathleen Perez-Lopez, and Emily Rosenthal. The study authors are also grateful to Rachel Garrett, Jenifer Harr-Robins, Luke Keele, and Paul Bailey for their help with data analyses. Finally, we are grateful to Connie Conroy who provided administrative assistance throughout the project.

This page has been left blank for double-sided copying.



## **Disclosure of Potential Conflicts of Interest**

The research team was comprised of staff from American Institutes for Research (AIR). None of the research team members has financial interests that could be affected by findings from the Early Implementation Findings From a Study of Teacher and Principal Performance Measurement and Feedback. No one on the 10-member technical working group, convened by the research team three times to provide advice and guidance, has financial interests that could be affected by findings from the evaluation.

This page has been left blank for double-sided copying.

# Executive Summary

Educator performance evaluation systems are a potential tool for improving student achievement.<sup>1</sup> By removing ineffective teachers and principals and/or through increasing the effectiveness of the existing workforce, such systems may result in higher student achievement.<sup>2</sup>

Emerging research suggests some promising features of performance evaluation measures. For example, research suggests that, to measure classroom practice, additional observations of the same teacher, beyond the first, form a more reliable measure of a teacher's typical practice, especially when more than one observer is used.<sup>3</sup> There is also some evidence from recent research that giving more frequent, specific feedback on classroom practice may lead to improvements in teacher performance and student achievement.<sup>4</sup>

The U.S. Department of Education's Institute of Education Sciences is conducting a study on the implementation and impacts of teacher and principal performance measures that are consistent with emerging research. As part of the study, eight districts were provided resources and support to implement the following three performance measures in a selected sample of schools in 2012-13 and 2013-14:

- a measure of teacher classroom practice with subsequent feedback sessions conducted four times per year, based on a classroom observation rubric;
- a measure of teacher contributions to student achievement growth (i.e., value-added scores), provided to teachers and their principals once per year; and
- a measure of principal leadership with subsequent feedback sessions conducted twice per year, using a leadership survey administered to the principal, the principal's supervisor, and the principal's teachers.<sup>5</sup>

No formal "stakes" were attached to the measures – for example, they were not used by the study districts for staffing decisions such as tenure or continued employment.<sup>6</sup> Instead, the measures were used to provide educators and their supervisors with information regarding performance.

---

<sup>1</sup> See Stecher et al. (2016); Weisburg, Daniel, Sexton, Mulhern, and Keeling, (2009).

<sup>2</sup> Researchers studying a range of educator workforce interventions theorize that there are two key mechanisms leading to improved student achievement: changes in the composition of the workforce and changes in the skills of continuing workers. See, for example, Taylor and Tyler (2012) regarding teacher evaluation, Chiang et al. (2015) regarding performance-based compensation, and Glazerman et al. (2010) regarding comprehensive teacher induction.

<sup>3</sup> See Bill & Melinda Gates Foundation (2012); Whitehurst, Chingos, and Lindquist (2014).

<sup>4</sup> See Steinberg and Sartain (In Press); Taylor and Tyler (2012).

<sup>5</sup> These measures are the kinds of measures emphasized through the Elementary and Secondary Education Act flexibility waivers, as well as federal programs such as Race to the Top and the Teacher Incentive Fund. They also are the kinds of measures that states are allowed to develop and support with Title II, Part A funds under the Every Student Succeeds Act of 2015.

<sup>6</sup> There were exceptions in three districts. In these districts, the observations conducted by principals as part of this study counted in their official rating system if the teacher was due to be observed that year under the district's existing evaluation system.

Such information might affect perceptions about performance, motivate improvement, and/or lead to improved knowledge and skills among educators.

This report focuses on the first year of the two years of implementation, describing the characteristics of the educator performance measures and teachers' and principals' experiences with performance feedback. It is one of the few detailed descriptions of educator performance measures implemented on a large scale in districts.<sup>7</sup> The final report will primarily focus on impacts on outcomes including principal leadership, teacher classroom practice, and student achievement but will also include descriptive information on the second year of implementation.

The main findings in this report are as follows:

- **Educator performance measures were implemented generally as planned, except that fewer than the intended number of educators accessed the student growth reports.** Most teachers received the intended four rounds of observations and feedback sessions (mean = 3.8 observations and 3.7 feedback sessions per teacher) by trained and certified observers. Only 40 percent of principals and 39 percent of teachers with value-added scores accessed their student growth reports. All principals in the implementing schools received two reports about their leadership practices based on their survey responses and that of their teachers and their supervisor. They also met with their supervisors to discuss their reports, in both fall and spring.
- **Both classroom observation and student growth measures differentiated teacher performance, but observation scores were skewed toward the upper end of the scale.** In each classroom observation window, a large majority of the teachers observed had classroom observation overall scores in the top two performance levels (more than 85 percent), and very few teachers had overall scores in the lower two levels (less than 15 percent). However, there was some variation in scores across teachers, and both value-added scores and average classroom observation scores over the year had sufficient reliability to capture performance differences among some teachers. About one quarter of the teachers with value-added scores in reading/English language arts and one half of the teachers with value-added scores in mathematics had student growth reports indicating that their score in that subject was measurably below or above the district average.<sup>8</sup>
- **The principal leadership measure differentiated performance, but there was limited consistency in scores across survey respondent groups.** Principals' overall leadership scores were spread across the four performance levels. Half or more of

---

<sup>7</sup> For an example, see Lipscomb, Terzиеv, and Chaplin (2015).

<sup>8</sup> To test whether teachers' value-added scores differed from the district average, we used an 80 percent confidence interval. There were two reasons for the use of an 80 percent rather than a 95 percent confidence interval more typical of statistical tests. First, the student growth report available to principals and teachers in the intervention schools included the score with an 80 percent confidence interval. As part of the student growth report training, educators were told to use this confidence interval to determine if their score differed from average performance in the district. Second, the value-added scores were used for informational purposes and not decisions with consequences for employment, for example. Given this, the 80 percent confidence interval was intended to balance the risks associated with mistakenly classifying average performers as above or below average with the risks associated with mistakenly classifying above or below average performers as average.

principals in the schools selected to implement the intervention received overall scores on the principal leadership measure in the lower two categories labeled *basic* or *below basic* (70 percent in the fall and 51 percent in the spring administration). However, the respondent groups (principals, teachers, and principal supervisors) often scored principals differently.

- **Both teachers and principals in schools selected to implement the intervention reported receiving more feedback than those in schools in the same districts selected to continue with business-as-usual.** For example, teachers in intervention schools reported more feedback sessions that were accompanied by a rating and written narrative (3.0 versus 0.7 sessions).<sup>9</sup> They also reported spending more total time in feedback sessions (80 minutes versus 18 minutes). Principals in intervention schools reported more instances of receiving feedback with ratings (1.0 versus 0.4) and spending more total time in feedback sessions (60 minutes versus 41 minutes).

## Study Overview

The purpose of this study is to describe teachers' and principals' experiences with the study's performance measures and feedback over two years, and to examine whether the information provided by the measures and feedback affected educator and student outcomes. The study has five research questions:

1. To what extent were the performance measures and feedback implemented as planned?
2. To what extent did the performance measures distinguish educator performance?
3. To what extent did educators' experiences with performance feedback differ for treatment and control schools?
4. Did the intervention have an impact on teacher classroom practice and principal leadership?
5. Did the intervention have an impact on student achievement?

This report addresses the first three questions, focusing on the first year of implementation. A subsequent report will address the first three questions focusing on the second year of implementation. That report will also address the last two questions.

## Study Design

The study examines the implementation and impacts of an intervention consisting of three performance measures with feedback for teachers and principals through an experimental design conducted in eight purposefully selected districts. We recruited districts that met the following criteria: (1) had at least 20 elementary and middle schools, (2) had data systems that were sufficient to support value-added analysis, and (3) had current performance measures and feedback that were less systematic and intensive than that implemented as part of the study. Recruited districts required fewer than four observations of teachers per year. In addition, the

---

<sup>9</sup> Throughout this executive summary, all treatment-control differences noted are statistically significant using two-tailed tests and applying a 5 percent critical value, unless noted otherwise.

districts' evaluation systems did not require the inclusion of student achievement information in teacher ratings. None of the recruited districts used a leadership measure similar to that used by the study.

The study used two different observation measures to make the study findings more broadly relevant than would be the case if only one measure was used. Four of the eight study districts used the Classroom Assessment and Scoring System (CLASS) and the other four study districts used Charlotte Danielson's Framework for Teaching (FFT). The observation rubrics were not randomly assigned but rather assigned based on district preference. Thus, differences in the study results in the CLASS and FFT districts cannot necessarily be attributed to the CLASS and FFT observation systems; differences could occur due to other district characteristics.

Each study district identified a set of regular elementary and middle schools that were willing to participate in the study. In these schools, the study focuses on the teachers of mathematics and reading/English language arts in grades 4-8, as well as the principals.<sup>10</sup> The schools were assigned by lottery to implement the three measures with feedback (the treatment group) or not (the control group). Both groups continued to implement their district's existing performance evaluations and measures, and the treatment group additionally implemented the study's performance measures with feedback. In total, 63 treatment schools and 64 control schools participated in the study.

Consistent with the recruitment criteria, the study districts are larger and more likely to be urban than the average U.S. district. The study schools were similar to schools in the national population in terms of enrollment and Title I status, but on average had a higher percentage of students who were minorities.

Data for this report came from multiple sources as described next.

**Data on the implementation of the intervention.** We documented attendance at orientation and training events related to the study's performance measures. Online system records maintained by the vendors of the measures were used for information on observer certification test pass rates, the frequency and timing of teacher observations and feedback sessions, and teachers' and principals' accessing of student growth reports. Surveys of observers hired by the study and interviews with district officials provided further information regarding the implementation of the observations and the district context, respectively.

**Data on measures of educator performance.** Data on measures of teacher classroom practice, student growth, and principal leadership were collected through the vendors' online systems.

**Data on educators' experiences with performance feedback.** In spring 2013, we surveyed both the principals and teachers in all treatment and control schools. These surveys

---

<sup>10</sup> Teachers of Kindergarten through grade 3 also participated in the study. This was done mainly to promote schoolwide engagement in the implementation of the classroom practice and principal leadership performance measures. These teachers are not included in the main study analyses, however, because student assessment data are not available in Kindergarten through grade 3.

collected information on the nature and frequency of performance information educators received and their perceptions of that information.

**Data on the characteristics of study participants.** To compare the characteristics of participants in the treatment and control groups, we collected data on school characteristics from the 2011–12 Common Core of Data and collected data on principals', teachers', and students' characteristics from district administrative records.

## ***Analyses***

To examine the implementation of the performance measures, we describe the extent to which study participants received the training on the measures, carried out the performance measurement activities, and received performance information and feedback as planned. We also examined the characteristics of the ratings teachers and principals received, including whether they distinguish between lower and higher performers. These analyses yielded the average rating scores, the percentage of ratings in each performance level, and the variation in the ratings across teachers and across principals. To assess whether the study's intervention led to differences in educators' experiences with performance measurement and feedback, we compared survey responses of teachers and principals in the treatment and control groups.

## **Detailed Summary of Findings**

The following section provides additional information about the extent to which each of the study measures was carried out as intended and whether the information from the measures distinguished between lower- and higher-performing educators and thus could be used to identify educators in need of support. These analyses pertain only to teachers and principals in the treatment schools. This section also highlights the extent to which educators' experiences with performance information differed between the treatment and control groups.

### ***The Classroom Practice Measure and Feedback***

The teacher classroom practice measure was based on four classroom observations during the school year. For each teacher, one observation was to be conducted by a school administrator and the other three by observers hired by the study. After each observation, the observer was expected to prepare a standard report with both ratings and narrative justification and to discuss the report with the teacher during a feedback session. Both teachers and their principal had access to the standard report.

### **How Many Observations Were Conducted and What Were Observers' Qualifications?**

- **Observers were trained and certified as planned.** Nearly all observers (92 percent for CLASS and 97 percent for FFT) completed all of the required training, which lasted three days for CLASS and four days for FFT. All observers passed the certification test, although it took multiple attempts to pass the test for half of the CLASS trainees and 17 percent of the FFT trainees.

- **The majority of teachers were observed the intended four times and received feedback.** The majority of teachers (73 percent for CLASS and 95 percent for FFT) received all four observations, and the majority of teachers (57 percent for CLASS and 94 percent for FFT) also received all four feedback sessions, as intended. On average, teachers received 3.8 observations (3.7 for CLASS and 3.9 for FFT) and 3.7 feedback sessions (3.5 for CLASS and 3.9 for FFT) during the first year of the study.

### **What Were the Characteristics of the Classroom Practice Performance Information Provided?**

- **CLASS reports provided separate scores for individual dimensions as well as the teacher's overall score and a sense of how their performance compared with others; FFT reports provided only separate scores for individual dimensions.** The CLASS reports included scores for 12 dimensions of teaching grouped into four teaching domains, as well as an overall score for the observation and a score for each domain (emotional support, classroom organization, instructional support, and student engagement). In addition, the CLASS reports included comparisons with the district average scores and the teacher's prior scores. The FFT reports provided scores for up to 10 dimensions of teaching grouped into two teaching domains (classroom environment and instruction). The FFT reports did not include an overall score, domain scores, scores from past observations, or district average scores.
- **Most of the CLASS observation reports identified at least one dimension of classroom practice to improve and illustrated it with an example from the observation, but less than a quarter of FFT reports did so.** The observers were required to write narrative text identifying at least one dimension of practice as a strength and one dimension for improvement. The majority of the observation reports (76 percent of CLASS reports and 71 percent of FFT reports) did so. In addition, three quarters of the CLASS reports supported the identified dimension(s) for improvement with at least one example from the observation, but less than a quarter (23 percent) of the FFT reports did so.<sup>11</sup>
- **For both CLASS and FFT, observation scores were concentrated at the upper end of the scale, limiting the degree of differentiation between lower- and higher-performing teachers.** Nearly all teachers had CLASS or FFT overall scores for a given observation window in the top two performance levels (more than 95 percent of the CLASS scores and more than 85 percent of the FFT scores). Only a small percentage of the teachers had scores consistent with the lowest two performance levels (under 5 percent for CLASS and under 15 percent for FFT depending on the observation window).<sup>12</sup> (See exhibits ES.1 and ES.2.) While most teachers had overall scores in the top two performance levels, many teachers had dimension-level scores at different performance levels (e.g., in the first window 61 percent of CLASS teachers and 69 percent of FFT teachers received scores at multiple performance levels).

<sup>11</sup> The findings reported here are based on an analysis of 160 randomly selected reports.

<sup>12</sup> Teachers observed using the FFT instrument did not receive an overall score or overall performance level for each observation window. For analytic purposes, the study's evaluation team calculated each teacher's average score in each observation window based on the 1 to 4 rating for each dimension of practice.



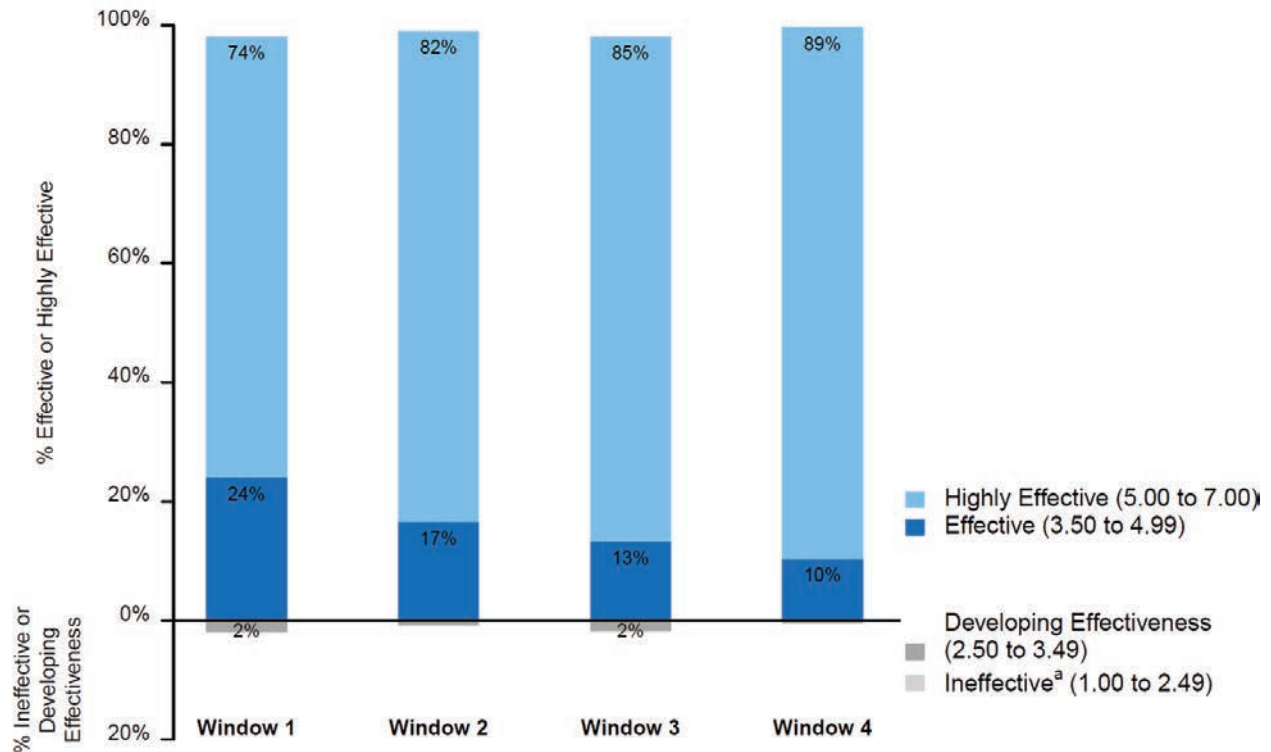
- **Teachers' overall classroom observation scores, averaged across all four windows, contained measurement error, but provided some reliable information to distinguish between lower- and higher-performing teachers and were positively correlated with teacher value-added scores.** Classroom observation scores averaged across the four observation windows had some reliability to help distinguish average teacher performance (reliability estimated between .42 and .50 for CLASS and .69 and .75 for FFT). These estimates, while lower than conventional thresholds for measures used in research, are consistent with findings from other studies of classroom observation reliability.<sup>13</sup> In addition, the CLASS and FFT four-window average scores were positively, although weakly, associated with teachers' prior-year value-added scores (correlations of .09 and .17, respectively).<sup>14</sup>
- **Differences in a teacher's ratings across observations limited how much one could learn about persistent performance from a single observation.** Less than half of the variation in teacher scores from a given observation window reflected stable classroom practice over the year. The reliability estimate for a single observation was .24 for CLASS scores and .49 for FFT scores, which indicates that 24 percent of the variation in CLASS scores and 49 percent of the variation in FFT scores reflected stable practice over the year.

---

<sup>13</sup> See Casabianca et al. (2013); Ho and Kane (2013); Kane and Staiger (2012).

<sup>14</sup> Although the correlations between classroom observation overall scores and value-added scores were modest in magnitude, these correlations are consistent with the magnitudes found by other studies (Chaplin et al. 2014; Kane and Staiger 2012; Kane et al. 2011) and likely underestimate the strength of the true association because of measurement error in both the observation scores and the value-added scores.

**Exhibit ES.1. Distribution of treatment teachers across performance levels based on CLASS overall scores, by observation window**



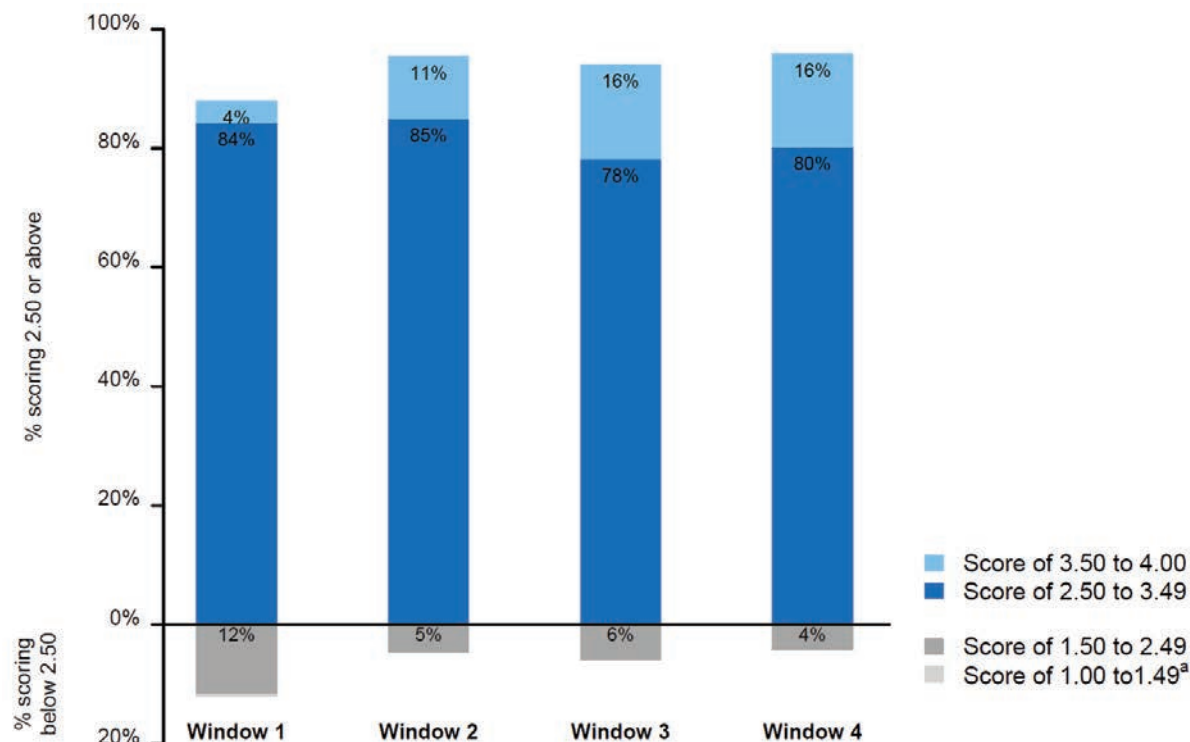
**Exhibit Reads:** Of treatment teachers in CLASS districts observed in window 1, 74 percent had a CLASS overall score at the *highly effective* performance level, 24 percent at the *effective* performance level, and 2 percent at the *developing effectiveness* performance level. Less than 1 percent of teachers had an overall score at the *ineffective* performance level.

NOTE: Performance level distributions are based on teachers' overall CLASS ratings in each window. Sample size = 262 teachers in window 1, 307 teachers in window 2, 309 teachers in window 3, and 272 teachers in window 4. Reported percentages may not sum to 100 percent because of rounding.

<sup>a</sup> Within a window, less than 1 percent of teachers had an overall score at the *ineffective* performance level.

SOURCE: Teachstone Online System.

**Exhibit ES.2. Distribution of treatment teachers across study-defined performance levels based on FFT overall scores, by observation window**



**Exhibit Reads:** Of treatment teachers in FFT districts observed in window 1, 4 percent had an FFT overall score between 3.50 and 4.00, 84 percent had a score between 2.50 and 3.49, and 12 percent had a score between 1.50 and 2.49. Less than one percent of teachers had an overall score below 1.50.

**NOTE:** The distribution in each window is based on teachers' FFT overall scores categorized into study-defined performance levels. To create the overall scores and performance levels, the study's evaluation team first calculated an overall score by averaging the teacher's ten FFT dimension scores, each of which was rated on a 1 to 4 scale. The overall scores were then categorized into study-defined performance levels by rounding them to the nearest whole number. This created four performance levels aligned with the FFT dimension scores. An FFT dimension score of 1 corresponds to *unsatisfactory*, 2 corresponds to *basic*, 3 corresponds to *proficient*, and 4 corresponds to *distinguished*. Average FFT scores and overall performance levels were not provided in the FFT reports teachers received. Sample size = 216 teachers in window 1, 219 teachers in window 2, 220 teachers in window 3, and 217 teachers in window 4. Reported percentages may not sum to 100 percent because of rounding.

<sup>a</sup> Within a window, less than 1 percent of teachers had an overall score below 1.50.

SOURCE: Teachscape Online System.

## The Student Growth Measure

The measure of student growth was designed to provide teachers with information on their contribution to student achievement, using value-added methods. Value added methods involve predicting the test score each student would have received, accounting for prior achievement and other characteristics, if the student had been taught by the average teacher in the district. A teacher's value added score is obtained by comparing the average actual performance of the teacher's students to the average of the students' predicted scores.

Teacher value-added scores were generated for all teachers of students in grades 4–8 reading/English language arts and mathematics in each district using the achievement data for

the students that each teacher taught in the previous two years.<sup>15</sup> Individual teachers in the treatment schools in these grades were given access to a report on their scores during the first year of implementation. Treatment principals were also given access to a report that included their teachers' student growth reports as well as school average value-added scores, overall and by subject and grade.

### Who Received the Student Growth Performance Information?

- **A large majority of teachers had sufficient data to produce student growth reports.** Overall, student achievement data were sufficient to compute value-added scores and produce student growth reports for 80 percent of the teachers, who were in grades 4-8.
- **Although most teachers and principals participated in the student growth report training, less than half of the teachers and principals accessed their reports.** Overall, 85 percent of teachers and 81 percent of principals participated in a webinar prior to the release of the student growth reports. The webinar oriented the participants to the value-added scores, the content of the student growth reports, and how to access them. The online reporting system showed that 40 percent of the teachers with value-added scores and 38 percent of the principals accessed their student growth reports.

### What Were the Characteristics of the Student Growth Performance Information Provided?

- **Student growth reports included school and individual teachers' value-added scores.** The teacher report included a teacher's overall and subject-specific value-added scores (both reading/English language arts and mathematics for those who taught both subjects) with an indication of their percentile ranking relative to other teachers in the district, and the average teacher score in the district and school. All scores included confidence intervals/standard error information to indicate the precision of the estimated scores. Each teacher could also access a roster that included the number and names of students used to calculate their score. For each teacher in his or her school, the principal could view an overall value-added score, scores by subject and grade, and scores across time. Principals could also view school average scores overall and by subject and grade.
- **Many teachers with a student growth report had a value-added score that measurably differed from the district average, particularly in mathematics.** The student growth reports available to teachers and principals included teachers' value-added scores along with an 80 percent confidence interval, which could be used to determine whether the scores were "measurably" different from the district's average teacher.<sup>16</sup> For example, in mathematics, 25 percent of the teachers had a value-added

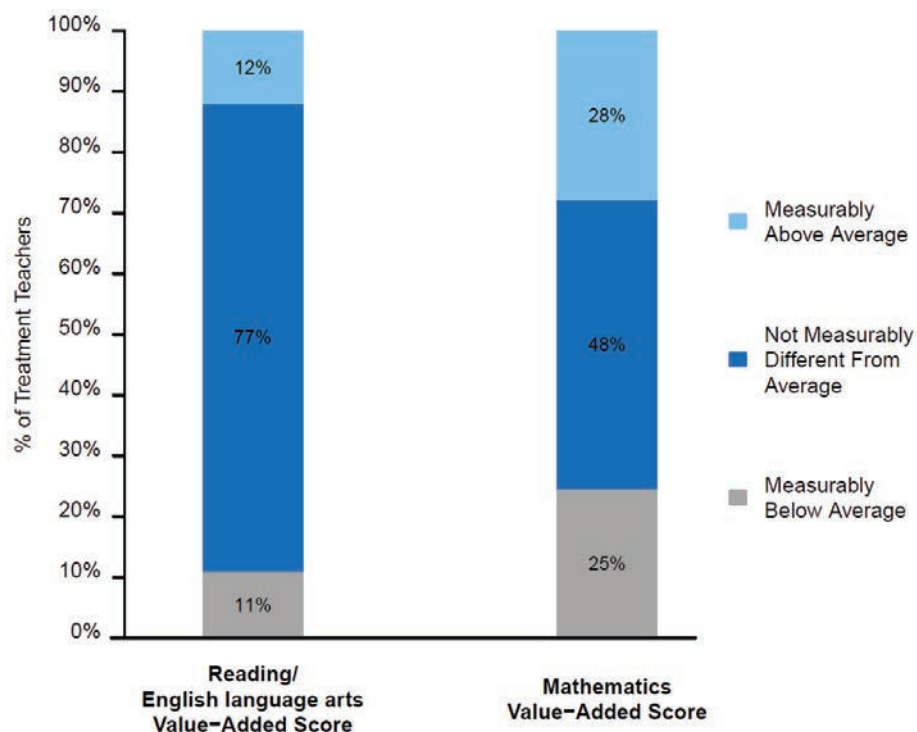
---

<sup>15</sup> A value-added score for a given subject was produced for a teacher only if the teacher had at least 10 students who had the necessary achievement data.

<sup>16</sup> The student growth reports used an 80 percent confidence interval (i.e., the range of scores that have an 80 percent chance of including the teacher's "true" score) to identify scores that were "measurably" below or above average. This benchmark was selected in order to appropriately balance the risk of misclassifying a teacher who is actually

score that was considered measurably below the district average, and 28 percent had a score that was considered measurably above average. See exhibit ES.3.

**Exhibit ES.3. Distribution of treatment teachers based on whether their value-added score was considered measurably above or below the district average, by subject**



**Exhibit Reads:** For treatment teachers with mathematics value-added scores, 28 percent had scores considered measurably above the district average.

**NOTE:** Distributions of teachers are based on whether the 80 percent confidence interval for a teacher's value-added score was above or below the district average. To indicate the amount of uncertainty around each teacher's score, the student growth reports included 80 percent confidence intervals, which showed the range of scores that have an 80 percent chance of including the teacher's "true" score. This benchmark was selected in order to appropriately balance two types of risks within the context of an intervention designed to provide feedback on performance without explicit consequences such as promotion or dismissal: (1) the risk of misidentifying truly average teachers as below- or above-average, and (2) the risk of misidentifying teachers who were truly below- or above-average as average teachers. Sample size = 338 teachers with mathematics value-added scores and 321 teachers with reading/English language arts value-added scores. Reported percentages may not sum to 100 percent because of rounding.

**SOURCE:** AIR value-added system.

## ***The Principal Leadership Measure and Feedback***

Feedback on principal leadership was based on the Vanderbilt Assessment of Leadership in Education (VAL-ED), a 360-degree survey assessment administered twice a year to principals, principal supervisors, and teachers. The VAL-ED includes six "core components" of principal performance: high standards for student learning, rigorous curriculum, quality instruction,

average as above or below average, against the risk of misclassifying a teacher who is actually above or below average as average. One consideration in striking this balance was that the study districts agreed that the value-added scores would not be used for decisions with consequences for employment. This reduced the potential downside associated with misidentifying an average teacher as below average.

culture of learning and professional behavior, connections to external communities and performance accountability. Principals are also rated on six “key processes”: planning, implementing, supporting, advocating, communicating, and monitoring. A report for each principal was generated after each administration of the VAL-ED, and the principal’s supervisor was expected to discuss the report with the principal in a feedback session.

### How Was the Principal Leadership Measure Implemented?

- **All principals and their supervisors received training on using VAL-ED.** All principals and their supervisors participated in a two-hour VAL-ED training in summer 2012. During the school year, all principals’ supervisors also received a one-hour training to prepare them to conduct the feedback sessions. In addition, teachers were offered a one-hour introduction to VAL-ED at the beginning of the school year, as well as an orientation webinar during the school year.
- **All VAL-ED reports incorporated input from the principal, the principal’s supervisor, and most teachers.** All principals and their supervisors completed the VAL-ED rating form, and a high percentage of teachers in each treatment school (80 percent in fall and 90 percent in spring on average) also completed the form.
- **All VAL-ED feedback sessions occurred as planned.** In both fall and spring, all principals met with their supervisors to discuss their VAL-ED reports. Principal supervisors reported feedback sessions lasting on average 52 minutes in the fall and 46 minutes in the spring.

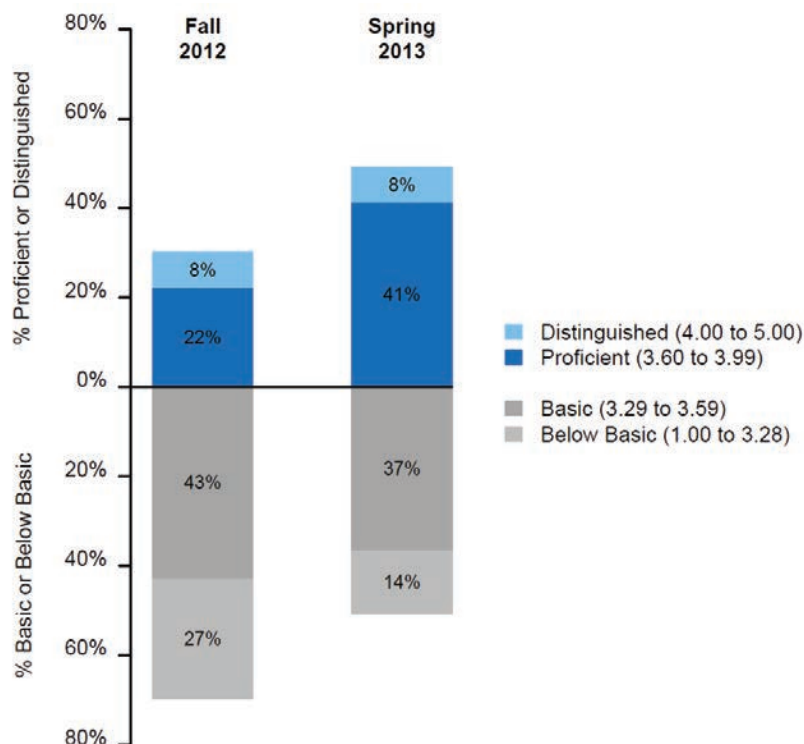
### What Were the Characteristics of the Principal Leadership Performance Information Provided?

- **The VAL-ED reports present scores and performance levels, as well as percentile ranks, for each dimension of leadership.** VAL-ED reports present an overall score, a score for each core component, and a score for each key process. For each of these 13 scores, the report additionally presents a performance label and a percentile rank, relative to the principals included in a national VAL-ED field test. Each score (i.e., overall score, core component scores, and key process scores) is an average across the three respondent groups (i.e., principal, supervisor, and teachers), with each group weighted equally. The report additionally shows the scores received from each respondent group separately.
- **The VAL-ED ratings classified some principals as lower-performing and some as higher-performing.** In the fall, principals’ overall scores were distributed across the four performance levels (8 percent of principals were labeled *distinguished*, 22 percent *proficient*, 43 percent *basic*, and 27 percent *below basic*). In the spring administration, half the principals received an overall score associated with a performance level of *proficient* or *distinguished* and half received a score at the *basic* or *below basic* level.<sup>17</sup> (See exhibit ES.4.)

---

<sup>17</sup> The increase in average VAL-ED overall scores from the fall to spring is primarily a product of an increase in the principal self-ratings. Average ratings of principal leadership based on the three respondent groups were similar in

**Exhibit ES.4. Distribution of treatment principals across performance levels based on VAL-ED overall scores, by assessment window**



**Exhibit Reads:** In fall 2012, 8 percent of treatment principals had a VAL-ED overall score at the *distinguished* performance level, 22 percent at the *proficient* level, 43 percent at the *basic* level, and 27 percent at the *below basic* level.

NOTE: Performance level distributions are based on principals' VAL-ED overall scores at each assessment window. The overall score is an average of the scores from the principal's supervisor, teachers, and the principal's own self-rated score, with each group weighted equally. Sample size = 63 principals for both fall 2012 and spring 2013. Reported percentages may not sum to 100 percent because of rounding.

SOURCE: Fall 2012 and Spring 2013 VAL-ED Surveys.

- VAL-ED ratings provided by principals, supervisors, and teachers in the fall were often too different to form a reliable measure, but the spring ratings were consistent enough to distinguish between some lower- and higher-performing principals.** To provide information about a principal's overall effectiveness, the VAL-ED scores should communicate a consistent (i.e., reliable) message about the principal's effectiveness across the three respondent groups (the principal, the principal's supervisor, the principal's teachers). Based on the literature on 360-degree surveys, we would expect correlations between respondent group scores between .25 and .35.<sup>18</sup> In the fall, however, agreement among the three respondent groups' overall scores was low, with correlations ranging from .06 to .27. In the spring, correlations were higher (between .26 and .38), and thus the reports provided a more consistent message

the fall; however, in the spring, principal self-ratings were higher on average (3.76) than the ratings from their supervisors (3.50, p-value of the difference <.05) and teachers (3.57, p-value of the difference <.05).

<sup>18</sup> For the VAL-ED correlations, see Porter et al. (2010). For the literature on 360-degree surveys, see Conway and Huffcutt (1997).

about a principal's effectiveness. Viewing and discussing the fall reports may have led principals and their supervisors to better align their ratings in the spring.

## ***Educators' Performance Evaluation Experiences***

The study's performance measures were intended to provide educators with performance information that was more frequent, systematic, and useful as a guide for professional growth than the information that they normally receive. To assess whether this occurred, we compared the treatment and control groups' responses on surveys administered in the spring. Teacher surveys were usually completed at the beginning of the last of the four observation windows. Principal surveys were completed prior to the spring VAL-ED feedback session, which usually occurred at the end of the school year.

### **What Were Teachers' Experiences?**

- **Treatment teachers reported receiving more feedback on both their classroom practice and their students' achievement growth than control teachers.** Treatment teachers reported receiving more feedback sessions with ratings and a written narrative than control teachers (3.0 versus 0.7 instances). The average treatment teacher also received a larger amount of oral feedback than the average control teacher (80 minutes versus 18 minutes). Furthermore, relative to control teachers, treatment teachers were more likely to report receiving value-added scores (45 percent versus 24 percent) and less likely to report receiving test scores for individual students or classroom average scores.<sup>19</sup>
- **Among those who reported receiving feedback, treatment teachers indicated somewhat more positive perceptions than control teachers about the information they received on their classroom practice but not about the information on their students' achievement.** Although most teachers in both treatment and control groups reported agreeing or strongly agreeing that the feedback on their classroom practice provided specific ideas about how to improve, treatment teachers were more likely to report so (87 percent versus 79 percent). Almost all teachers (approximately 92 percent) in both groups indicated that the feedback on classroom practice was a fair assessment of their performance. Control teachers were more likely than treatment teachers to report that the student achievement information they received was easy to understand (89 percent versus 78 percent). However, less than half of the teachers in both groups agreed or strongly agreed that the achievement information was a fair assessment of their performance (49 percent for treatment teachers and 43 percent for control teachers, not a statistically significant difference) or a fair indicator of teacher effectiveness for all teachers (40 percent for treatment teachers versus 29 percent for control teachers, a statistically significant difference).

---

<sup>19</sup> This finding should be interpreted with caution because some teachers may not have had a correct understanding of the term "value-added scores." As a validity check, we compared treatment teachers' responses with electronic records indicating which teachers had accessed their own value-added scores in the online system, and we found that 34 percent of the treatment teachers who reported receiving value-added scores did not access their student growth reports in the online system, and 17 percent of treatment teachers who reported not receiving value-added scores actually accessed their online student growth reports.



## What Were Principals' Experiences?

- **Treatment principals reported receiving more feedback than control principals.** Treatment principals reported receiving feedback more often than control principals (2.0 versus 1.4 instances) and more instances of oral feedback with ratings (1.0 versus 0.4 instances). The average treatment principal also received more oral feedback than the average control principal (60 minutes versus 41 minutes). However, treatment principals were no more likely than control principals to report that their supervisors' feedback focused on specific topics related to VAL-ED.
- **Among those who reported receiving feedback, most principals in both treatment and control schools had positive perceptions about the feedback they received.** The majority (more than 70 percent) of the principals in both treatment and control schools agreed that the feedback they received was a fair assessment of their performance, and approximately two thirds or more of the principals agreed that the feedback they received contained specific ideas for improving their performance. Among those who received feedback, there was no statistically significant difference between treatment and control principals in their perceptions of the feedback.

## Future Report

This report focuses on findings from the first year of implementation of the study's three performance measures with feedback. Findings about the second year of implementation will be presented in the second-year study report. The second-year report also will present findings on the impact of the study's performance measures and feedback on teacher classroom practice, principal leadership, and student achievement.

